Categorical encoding of decision variables in orbitofrontal cortex

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Categorical encoding of decision variables in orbitofrontal cortex

Arno Onken[1,2]*, **Jue Xie**[3], **Stefano Panzeri**[1], **Camillo Padoa-Schioppa**[3]

1 Center for Neuroscience and Cognitive Systems, Istituto Italiano di Tecnologia, Rovereto, Italy, **2** School of Informatics, University of Edinburgh, Edinburgh, United Kingdom, **3** Department of Neuroscience, Washington University in St Louis, St Louis, Missouri, United States of America

* aonken@inf.ed.ac.uk

OPEN ACCESS

Citation: Onken A, Xie J, Panzeri S, Padoa-Schioppa C (2019) Categorical encoding of decision variables in orbitofrontal cortex. PLoS Comput Biol 15(10): e1006667. https://doi.org/10.1371/journal.pcbi.1006667

Editor: Francesco P. Battaglia, Radboud Universiteit Nijmegen, NETHERLANDS

Received: November 22, 2018

Accepted: September 2, 2019

Published: October 14, 2019

Data Availability Statement: Data and code are available from figshare: Onken, Arno; Xie, Jue; Panzeri, Stefano; Padoa-Schioppa, Camillo (2019): Categorical encoding of decision variables in orbitofrontal cortex. figshare. Dataset. https://doi.org/10.6084/m9.figshare.9844349.v1

Funding: This work was supported by the EU Marie Curie H2020-MSCA-IF-2014 Programme (https://ec.europa.eu, Grant No 659227 to AO), by the BRAIN Initiative (https://www.braininitiative.nih.gov, Grant No. R01-NS108410 to SP), and by the National Institutes of Health (https://www.nih.gov,

Abstract

A fundamental and recurrent question in systems neuroscience is that of assessing what variables are encoded by a given population of neurons. Such assessments are often challenging because neurons in one brain area may encode multiple variables, and because neuronal representations might be categorical or non-categorical. These issues are particularly pertinent to the representation of decision variables in the orbitofrontal cortex (OFC)–an area implicated in economic choices. Here we present a new algorithm to assess whether a neuronal representation is categorical or non-categorical, and to identify the encoded variables if the representation is indeed categorical. The algorithm is based on two clustering procedures, one variable-independent and the other variable-based. The two partitions are then compared through adjusted mutual information. The present algorithm overcomes limitations of previous approaches and is widely applicable. We tested the algorithm on synthetic data and then used it to examine neuronal data recorded in the primate OFC during economic decisions. Confirming previous assessments, we found the neuronal representation in OFC to be categorical in nature. We also found that neurons in this area encode the value of individual offers, the binary choice outcome and the chosen value. In other words, during economic choice, neurons in the primate OFC encode decision variables in a categorical way.

Author summary

Mental functions such as sensory perception or decision making ultimately rely on the activity of neuronal populations in different brain regions. Much research in neuroscience is devoted to understanding how different groups of neurons support specific brain functions by representing behaviorally relevant variables. In this respect, one important question is whether neuronal populations represent discrete sets of variables (categorical encoding) or random combinations of variables (non-categorical encoding). Here we developed a new algorithm to assess this general issue. We then used the algorithm to examine neurons in the orbitofrontal cortex (OFC) recorded while non-human primates performed economic decisions. We found that the neuronal representation was categorical. Specifically, neurons in the OFC encoded the value of individual offers, the binary

Grant No R01-MH104494 to CPS). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

choice outcome, and the chosen value. The present results support the hypothesis that economic decisions are formed within the OFC.

Introduction

A recurrent question in systems neuroscience is that of understanding what variables are encoded by a given population of neurons. Addressing this issue is a prerequisite to understand what role neurons play in functions such as sensation or decision making. In a typical experiment, animal subjects perform some task, behavioral conditions vary along one or more dimensions, and the corresponding parameter(s) define variables potentially encoded by neurons in some brain area. In first approximation, if firing rates vary systematically with a variable, it can be said that neurons encode or represent that variable. Building on this concept, countless studies shed light on the neural substrates of sensory, associative and motor processes. Importantly, identifying the variables encoded by a given population can sometimes be challenging due to the trial-by-trial variability of neuronal firing rates combined with three other factors. First, different neurons, even in close proximity to one another, may encode different variables, and the number of variables encoded by a neuronal population is generally not known. This situation may arise in any brain area but is most typical for prefrontal regions. Second, different candidate variables potentially encoded by the neuronal population may be substantially correlated with one another. Third, the encoding of different variables may be categorical or non-categorical. In a categorical representation, neurons in a population encode a discrete set of variables. Conversely, neurons in a category-free representation encode a continuum of variables [1–3]. Of course, the encoding scheme adopted by any particular population is not known a priori.

All these issues are particularly pertinent to the representation of decision variables in the orbitofrontal cortex (OFC)–an area implicated in economic (or value-based) decisions [4, 5]. In recent years, numerous studies have shown compelling evidence for mixed selectivity and category-free encoding in lateral prefrontal regions [6–10], suggesting that these traits are the hallmark of neural systems supporting complex cognitive functions [3, 11]. At the same time, several studies argued for categorical encoding of decision variables in the OFC. Concurrent results in this sense came from studies of economic decisions in non-human primates [12, 13] and from studies of decision confidence in rodents [14]. In contrast with these observations, a recent study argued for non-categorical encoding of decision variables in the primate OFC [15] (more on this below). Importantly, the categorical nature of this representation is a key assumption underlying current neuro-computational models of economic decisions [16–23]. Given the importance of this matter, we set up to revisit the question of categorical versus category-free encoding in the OFC using a new and more powerful statistical approach. Our goal was to develop a set of procedures (or algorithm) with four objectives in mind. First, the algorithm should assess the categorical versus category-free nature of a neuronal representation without committing to any particular set of variables. Second, if the encoding was indeed categorical, the algorithm should facilitate a quantitative comparison of multiple candidate variables potentially represented by the neuronal population. Third, the algorithm should operate seamlessly in cases where different variables encoded in the neuronal population are correlated. Fourth, the algorithm should be amenable to general use, for any neuronal population and any behavioral task.

To achieve our stated objectives, we considered the high-dimensional space defined by all the behavioral conditions occurring in the task (referred to as "trial types"). We noted that

each neuronal response corresponds to one point in this space. Furthermore, after normalization, each response corresponds to one point on the hyper-spherical surface of unitary radius. Cast in this terms, the problem of assessing whether a neuronal representation is categorical in nature maps onto a clustering problem defined on a high-dimensional hyper-spherical surface, which we resolve using a spherical k-means approach [24]. In our algorithm, the categorical or non-categorical nature of the representation is assessed before defining any behavioral variable. The spherical k-means returns a number of clusters and their locations in the space of possible responses (i.e., the hyper-spherical surface). Furthermore, any variable possibly encoded in the neuronal population (i.e., any quantity systematically varied across behavioral conditions) also corresponds to a point on the hyper-spherical surface. Casting a wide net, we can generate a large number of variables potentially encoded by the neuronal population and thus identify the subset of variables that minimizes the total distance from the clusters. Importantly, these procedures are completely general and do not depend on the specifics of the behavioral task, except for the definition of candidate variables potentially encoded by the neuronal population.

The Results are organized as follows. The first section describes the juice choice experiments conducted in monkeys, the neuronal data set collected in OFC, and previous analyses of these data. The second section introduces the new algorithm. The third section demonstrates how the criteria previously used to assess the categorical nature of the neuronal representation in OFC [12] can, in some cases, lead to erroneous conclusions. The fourth section describes the results obtained by testing the new algorithm on a set of synthetic data. In the following section, we describe the results obtained by analyzing the actual OFC data with the new procedures [25]. In a nutshell, the results corroborate previous findings [13]. In the Discussion, we compare the present algorithm to other approaches proposed in the literature. We also emphasize that procedures presented here provide a general and powerful method to analyze heterogeneous populations of neurons.

Results

Data set and previous analysis

In the experiments, two rhesus monkeys performed an economic choice task [13, 25]. In each session, the animal chose between two juices offered in variable amounts. The preferred and non-preferred juices were labeled juice A and juice B, respectively. A "trial type" was defined by two offers and a choice (e.g., [1A:3B, B]). The number of trial types varied from session to session (because we varied offer types and because of variability in choices), and each session typically included 5–20 trials per trial type. Our data set included 1008 neurons. Neuronal spiking activity was recorded and processed with standard techniques (see Methods). For the analysis of how firing rates depended on the task variables, we defined several time windows aligned with respect to different behavioral events. For each trial type and each time window, firing rates were averaged across trials. A "neuronal response" was defined as the activity of one cell in one time window as a function of the trial type.

Our previous analyses proceeded as follows [13, 25]. First, each neuronal response was tested with an ANOVA (factor trial type). Responses that passed a statistical criterion ($p<0.001$) were considered task-related and analyzed further. Our data set included 2047 task-related responses. Second, we defined a large number of variables potentially encoded by this population. We performed a linear regression of each response on each variable, from which we obtained the regression slope and the R^2. If the regression slope was significantly different from zero ($p<0.05$), the variable was said to "explain" the response. Third, two procedures–stepwise and best subset–were used to identify a small set of variables that best explained the

neuronal population. In a first study [13], both procedures identified variables *offer value A*, *offer value B*, *chosen value* and *chosen juice*. This result was replicated several times, including in the data set examined here [25] (**Fig 1**). Finally, each neuronal response was assigned to the selected variable that provided the highest R^2. Two additional analyses were conducted to address the issue of categorical versus non-categorical encoding. First, for each neuronal response it was assessed whether adding a second variable to the regression (through a bi-linear regression) would significantly improve the fit. This analysis found that this was the case for only a small fraction of responses [13]. A second analysis quantified for each neuronal response and for each pair of selected variables the difference in the corresponding R^2 (ΔR^2), and examined the distributions of ΔR^2 across the neuronal population. In general, these distributions presented a significant dip close to zero, indicating that variables were encoded in a categorical way [12].

The approach for data analysis summarized above has the advantage that it allows to examine a large number of variables in parallel without biasing the conclusions, and that it withstands situations in which candidate variables are highly correlated with one another [13]. At the same time, this approach presents two limitations. First, the analyses require to first define candidate variables, then identify the most explanatory ones, and finally assess whether the encoding is categorical. In contrast, it would be preferable to assess whether the encoding is categorical without committing to any particular variable or set of variables, and only later define variables that best capture each category of responses. Second, there are situations in which the argument for categorical encoding based on the distribution of ΔR^2 is not valid (more on this below). The algorithm presented in this study addresses these limitations.

Detection of categorical encoding using spherical clustering

The algorithm used to assess categorical encoding was applied to task-related responses (i.e., responses that passed the ANOVA criterion; see above). To detect categorical encoding, we devised an algorithm that combines a clustering procedure partitioning neural responses based only on their spatial configuration with one that starts from a particular set of variables. In essence, the idea is to select a set of variables that best represents the spatial configuration of neural responses in the high-dimensional space of trial types.

Fig 2 illustrates the algorithm for a 3-dimensional space (i.e., 3 trial types). Each data point represents a neuronal response (i.e., the activity of one cell, in one time window, averaged across trials for each trial type). Neuronal responses are first centered and normalized. This transformation places neuronal responses on a spherical surface of unitary radius. This data set undergoes two separate procedures for spherical clustering. First, data are examined with spherical k-means procedure, which does not assume any particular variable and yields a partition of the neural activity points based solely on the configuration of points in the high-dimensional space of trial types. For any number of clusters, this procedure alone reveals the categorical or non-categorical nature of the neuronal representation. Second, we perform a variable-centroid clustering, which starts from a particular subset of variables (iteratively chosen from a large set of candidate variables; see **Table 1**). Notably, each variable corresponds to a point on the spherical surface. Thus the subset of variables defines a corresponding number of cluster centroids, and we assign each neural response to the closest centroid. Each of these two clustering procedures (spherical k-means and variable centroid clustering) returns a partition of the population of neural responses. Importantly, the number of clusters is not known a priori. Furthermore, for any such number, there are many possible subsets of variables. We thus want to identify the subset of variables that best describes the neuronal data. As a measure of similarity between the two partitions, we use the adjusted mutual information [26]. Thus,

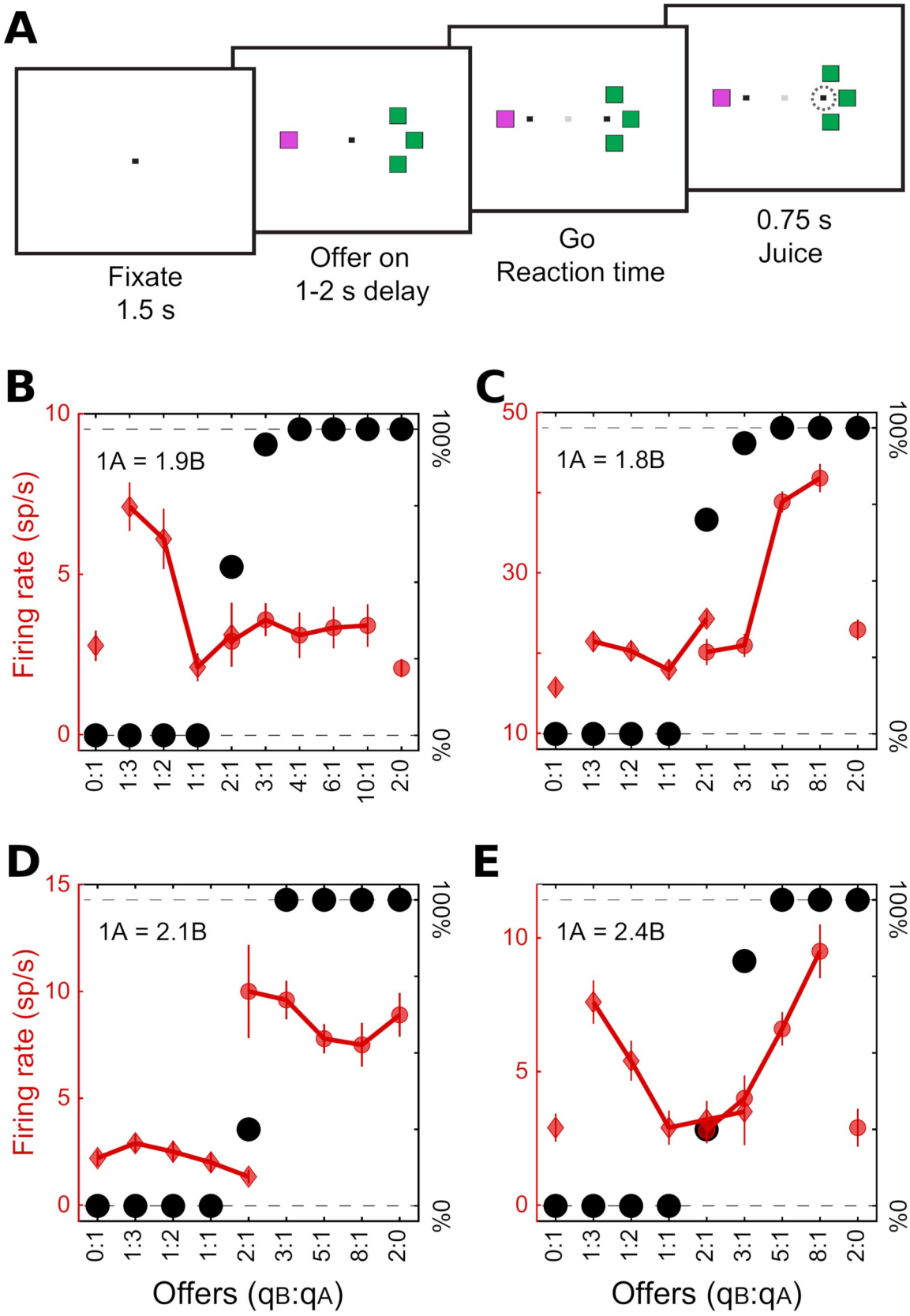
A
Fixate
1.5 s
Offer on
1-2 s delay
Go
Reaction time
0.75 s
Juice
B
1A = 1.9B
Firing rate (sp/s)
100%
%0
0:1
1:3
1:2
1:1
2:1
3:1
4:1
6:1
10:1
2:0
C
1A = 1.8B
5:1
8:1
D
1A = 2.1B
E
1A = 2.4B
Offers (qB:qA)

Fig 1. Task design and cell types. (A) Task design. In the experiments, monkeys chose between different juices offered in variable amounts. The two juices were labeled A and B, with A preferred. Offers were presented as visual stimuli on a computer monitor. Different juice types were associated with different colors, and the number of squares represented the juice quantity. After a randomly variable delay, the animal indicated its choice with an eye movement. (B) Example offer value A response. In this panel, the x-axis represents different offer types ranked by the ratio q_B/q_A, where q_J is the quantity of juice J offered. For each offer type, a black dot indicates the percent of trials in which the animal chose juice B (y-axis on the right). The relative value of the two juices (ρ) was obtained from a logistic fit. For this session, we measured ρ = 1.9. Red symbols illustrate one neuronal response. Diamonds and circle refer to trials in which the animal chose juice A and juice B, respectively. Vertical error bars indicate SEM. The activity of this cell increased almost linearly with the quantity of juice A offered, and did not depend on the quantity of juice B offered. (C) Example offer value B response. In this case, the response increased with the quantity of juice B offered, independently of juice A. (D) Example chosen juice B response. This response was nearly binary–high when the animal chose juice B and low when the animal chose juice A, independently of the quantity. (E) Example chosen value response. This response increased with the value of the chosen option, independently of the chosen juice. For chosen juice response, negative encoding for one juice is indistinguishable from positive encoding of the other juice (with this task design). Conventions in panels (C)-(E) are as in panel (B).

https://doi.org/10.1371/journal.pcbi.1006667.g001

we repeat the spherical k-means and the variable centroid clustering procedures for various number clusters and subset of variables. The variables that best match the non-committed spherical k-means partition are identified as encoded by the neuronal population.

Limits of previous approaches

In previous work, the categorical nature of the neuronal representation in OFC was assessed through the analysis of the distribution of ΔR^2 [12]. As explained above in Section "Data set and previous analysis", ΔR^2 quantifies the difference between the R^2 values obtained from the linear regressions onto a pair of selected variables. Intuitively, clusters of neural responses around variables should be discernible as peaks in the distribution of ΔR^2 values across the neuronal population. A dip in this distribution was thus interpreted as evidence for categorical encoding whereas a distributions without a dip was interpreted as evidence against categorical encoding [12]. We will now demonstrate that this criterion can sometimes lead to erroneous conclusions. To do so, we construct two synthetic data sets and we show that the ΔR^2 analysis fails while the spherical k-means algorithm reveals the true nature of the data. Again, each

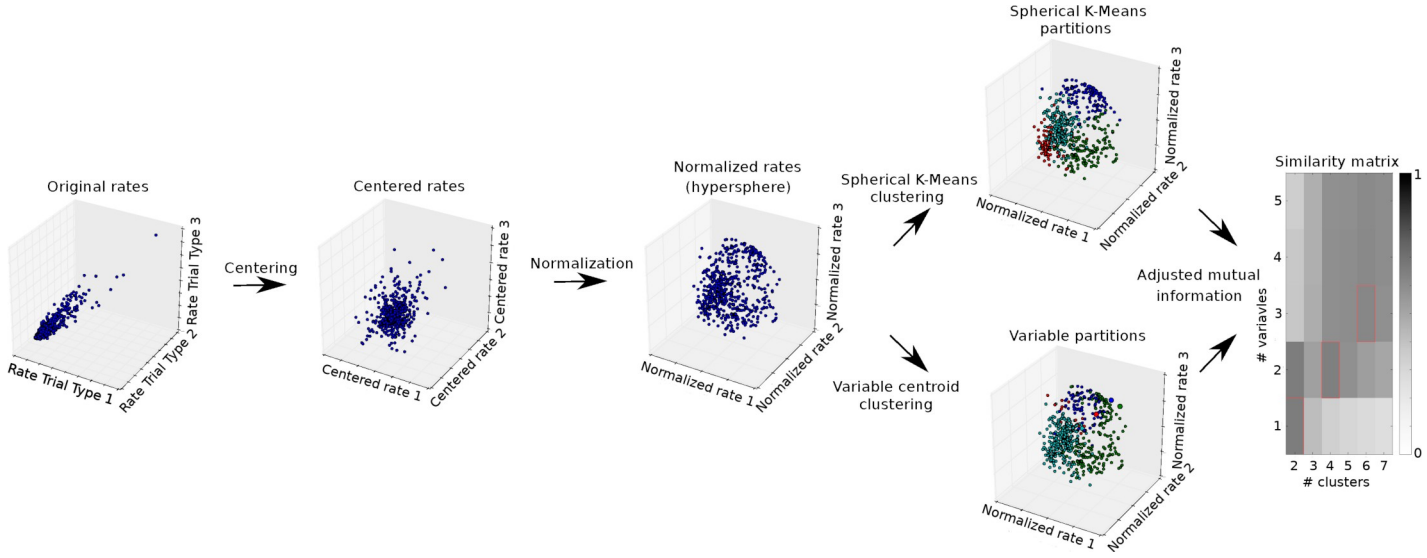

Fig 2. Illustration of the procedure for category discovery. For the original rates (left), each axis of the space denotes the firing rate of the cells in the various trial types. Each data point in this space represents a cell. For illustration, we present only 3 of the 9 dimensions. The original rates are then centered and normalized to unit length. The normalization effectively moves the points to the surface of a hyper-sphere. The points are then clustered using spherical k-means for a given number of clusters and centroid clustering for given variables. In the illustration, variables are represented as larger points. The resulting partitions are compared using the adjusted mutual information measure as a function of the number of clusters and number of variables (right).

https://doi.org/10.1371/journal.pcbi.1006667.g002

Table 1. Considered list of variables that are potentially encoded by the population.

No	Variable Name	Description
1	*offer value A*	Value of juice A offered
2	*offer value B*	Value of juice B offered
3	*chosen value*	Value of the chosen juice
4	*chosen value A*	Value of juice A chosen
5	*chosen value B*	Value of juice B chosen
6	*other value*	Value of the non-chosen juice
7	*value difference*	*chosen value—other value*
8	*value ratio*	*other value / chosen value*
9	*chosen number*	Chosen number
10	*chosen juice*	1 if A chosen, 0 if B chosen

https://doi.org/10.1371/journal.pcbi.1006667.t001

neural response is a point on the hyper-spherical surface of a high-dimensional space defined by the trial types, and variables correspond to points on this surface.

To simulate different neuronal populations, we generated distributions of points on the hyper-spherical surface.

Fig 3A illustrates the first example. Here data points form two clusters: a circular cluster close to the spherical pole and a banana-like cluster along the equator. Importantly, the distribution used to generate the banana cluster was uniform on a banana domain (no intrinsic dip). We now examine the situation in which the analyst identified the wrong variables, shown as large circles in Fig 3A. We assume that the analyst correctly identified the pole variable, but erroneously selected two variables located at the opposite tips of the banana cluster. As illustrated in Fig 3B, the distribution of ΔR^2 between the two banana variables has a significant dip around zero (Hartigan's dip test, $p<0.001$) suggesting that the two variables are categorically distinct. However, this suggestion is at odds with the ground truth. The dip in the distribution of ΔR^2 is due to the presence of the third cluster, because some of the data points in the banana are closer to the pole variable than to either of the banana variables. Hence, a dip in the distribution of ΔR^2 does not necessarily imply that the corresponding variables are encoded by categorically distinct groups of neurons. Importantly, the spherical k-means clustering correctly identifies the presence of two clusters (colors illustrate the k-means partitioning).

The second example makes the converse point, namely that a unimodal or uniform distribution of ΔR^2 does not necessarily imply a non-categorical representation. One obvious reason why this is the case is that absence of evidence is not evidence of absence; here we illustrate a subtler issue. We consider the same clusters defined above. In this case, we assume that the analyst correctly identified two variables, one in the pole cluster and one in the banana cluster. However, we assume that the variable in the banana cluster is off center (Fig 3C). As illustrated in Fig 3D, the resulting ΔR^2 histogram does not present a dip (Hartigan's dip test, $p = 0.313$), even though the two clusters are categorically separated. Importantly, the spherical k-means clustering correctly identifies the two clusters.

In conclusion, a dip in the distribution of ΔR^2 is neither sufficient nor necessary to assess the categorical nature of a neuronal representation. In general, such assessment requires the examination of the spatial distribution of data points in a high-dimensional space, using an approach such as the spherical k-means clustering.

Analysis of synthetic data

We considered several clustering procedures, and wanted to validate our algorithm to assess the categorical versus non-categorical nature of a neuronal representation on data for which

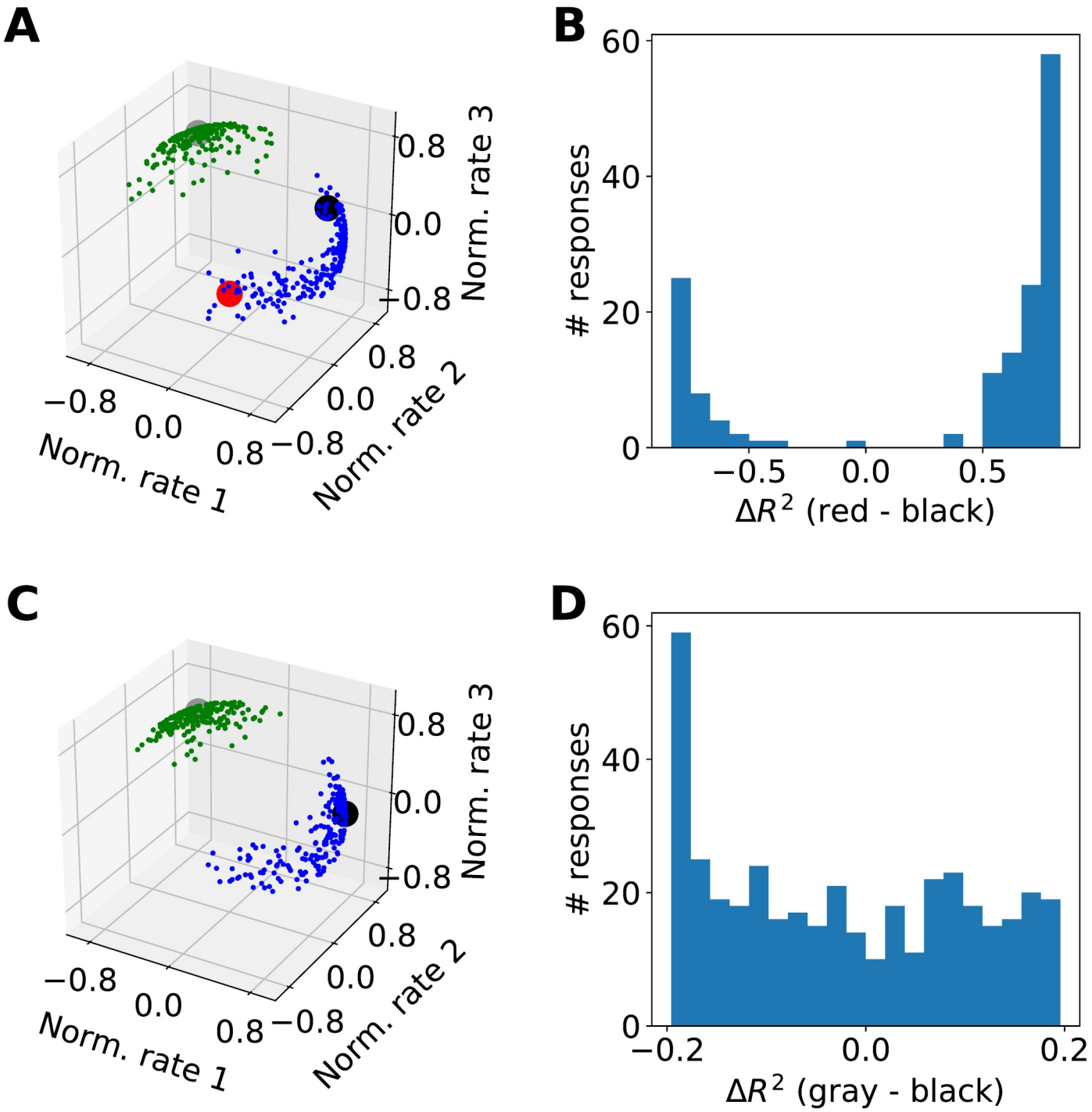

Fig 3. Two examples of how the ΔR^2 metrics can fail. (A,B) A dip in the distribution of ΔR^2 does not necessarily imply categorical encoding. The clustering algorithm yields two clusters. However, the analyst might have erroneously concluded that there are three variables, including two variables located in the tips of the banana cloud (red and black). The dip in the ΔR^2 histogram suggests that these two variables are encoded by categorically distinct populations, but this is in fact not the case. (C,D) Categorical encoding does not always result in a dip in the distribution of ΔR^2. In this case, we assume that the analyst correctly concluded that there are two variables, but might have defined these variables such that one is on the north pole (gray) and the other is on the east end of the banana cluster (black). Inspection of the ΔR^2 histogram does not reveal any dip. The reason is that data points on the west end of the banana cluster are equally far from the two variables.

https://doi.org/10.1371/journal.pcbi.1006667.g003

we could control the ground truth. Thus, we generated synthetic populations of neural responses with and without specific categorical structure, and applied clustering algorithms to these synthetic data.

For the real data, the experiments included 9 or 10 trial types, resulting in 9- or 10-dimensional neuronal responses, represented as points on the unitary hyper-sphere in 9 or 10 dimensions. (see section *Data set and previous analysis*). To generate synthetic neuronal responses with categorical nature, we randomly generated 9-dimensional points on the hyper-spherical surface clustered in the vicinity of selected variables (see Methods). We then analyzed these synthetic data sets with a wide range of clustering algorithms, including centroid-based clustering methods (mini-batch k-means, spherical k-means), hierarchical clustering methods (Ward, agglomerative clustering, Birch), and a graph-based clustering method (spectral clustering) [27–32]. To estimate the performance of these algorithms, we used silhouette plots, which are a common method to assess the goodness of clustering partitions [33]. For each data point X (here X is a normalized neuronal response), the silhouette value quantifies the mean distance between X and other data points in the same cluster, and compares it to the mean distance between X and data points in the nearest other cluster. The greater the silhouette value, the better the clustering. A negative silhouette value indicates that X was assigned to the wrong cluster, since X is closer to the nearest other cluster.

Fig 4 shows the silhouette plots obtained for the various clustering algorithms. We found that the hierarchical clustering methods (Ward, Agglomerative, Birch) produced the greatest number of negative silhouette values. Spectral clustering produced slightly less negative silhouette values than Ward as the best hierarchical clustering method. The centroid-based methods had no (spherical k-means) or very few (mini-batch k-means) negative silhouette values and many large silhouette values, suggesting that these methods found the most consistent clustering partitions. The silhouette analysis further suggests that the spherical k-means clustering is best suited for categorical data lying on a hyper-sphere. We also compared the silhouette plots on real data recorded from OFC. We found that spherical k-means had the smallest number of negative silhouette values, confirming the results from synthetic data. Of note, the superior performance of spherical k-means might be due to the fact that this algorithm explicitly considers the hyper-spherical structure of the data. Hence, we used spherical k-means for clustering in the remainder of this study.

We next compared the spherical k-means silhouette plots for categorical synthetic data with those for non-categorical data (**Fig 5**). To simulate neural responses without specific

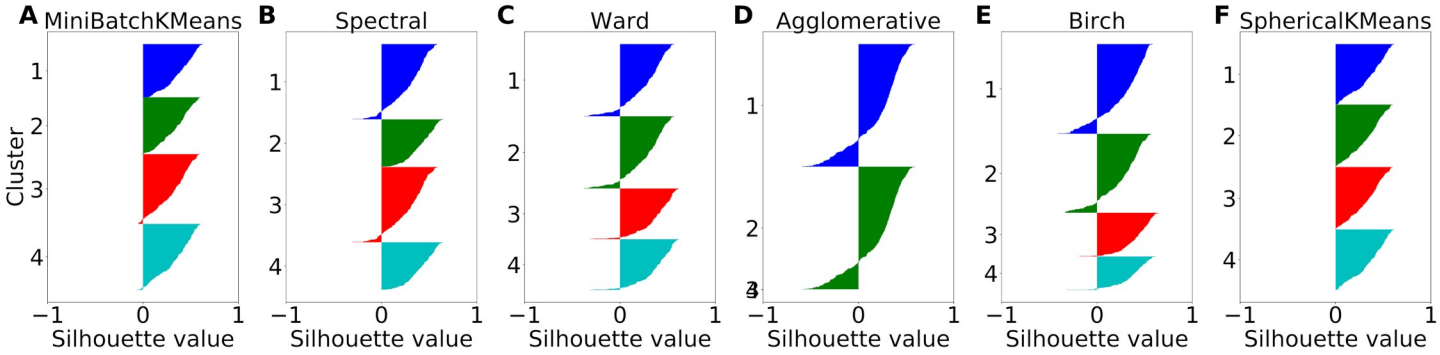

Fig 4. Silhouette comparison of clustering algorithms on synthetic categorical data. Synthetic data consist of firing rates from a total of 400 simulated cells representing the variables chosen value, offer value A, offer value B and chosen juice (100 cells each). Independent Gaussian noise with a standard deviation of 0.25 and a mean given by the variable rates was used to simulate the activity of a cell. Each color corresponds to one cluster. Clustering algorithms were Mini-Batch k-means (A), Spectral Clustering (B), Ward (C), Agglomerative Clustering (D), Birch (E) and Spherical k-means (F).The number of clusters was fixed to 4.

https://doi.org/10.1371/journal.pcbi.1006667.g004

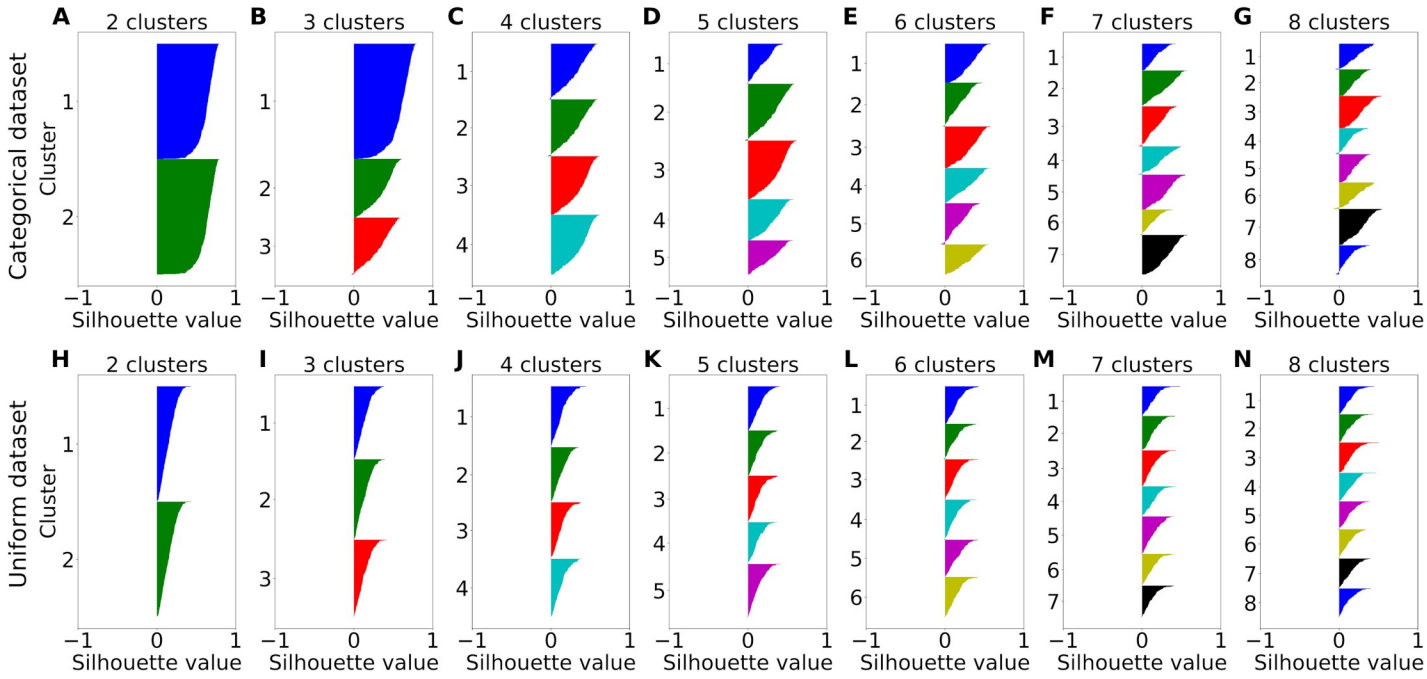

Fig 5. Silhouette comparison of spherical k-means clustering for different numbers of clusters on synthetic data. Synthetic data were either categorical (top row) simulated like in Fig 3 or non-categorical consisting of 400 samples uniformly distributed over the unit hyper-sphere (bottom row). For each data set, the number of clusters was varied between 2 and 8 (A-G) and (H-N) respectively. Each color corresponds to one cluster.

https://doi.org/10.1371/journal.pcbi.1006667.g005

categorical structure, we generated points uniformly on the hyper-spherical surface. We then varied the number of clusters between 2 and 7. We did not expect to find negative silhouette values for these data, because negative values indicate data point assignments to wrong clusters. Such mis-assignments cannot occur without any cluster structure in the data. Indeed, we did not find any negative silhouette values, neither for categorical data (**Fig 5A–5F**) nor for non-categorical data (**Fig 5G–5L**). However, while for categorical data the silhouette values in each cluster were dominated by large values yielding convex plots, the silhouette values for non-categorical data were dominated by small positive values yielding concave plots. Such concavity clearly indicates lack of cluster structure and allow to discriminate between categorical data and non-categorical data [33].

While the silhouette analysis provides a simple way to evaluate the assignments of data points to clusters, it does not immediately associate particular variables with clusters. To establish this relation, we devised a comparative clustering method. In addition to spherical k-means, we performed a centroid-based clustering where the centroids were defined by a particular set of variables. We refer to this procedure as "variable-centroid clustering". We assigned each data point to the nearest centroid on the sphere (see Methods). We then compared the clusters obtained from spherical k-means to the clusters obtained from variable-centroid clustering, and quantified their similarity for different sets of variables.

Quantifying the similarity of two clustering partitions is non-trivial because similarity should be invariant for cluster relabeling. Many measures of similarity have been proposed [26, 34, 35]. Here we tested three measures of similarity based on mutual information, which are founded on information theory and naturally satisfy our desiderata. Specifically, we tested mutual information (MI), normalized mutual information (NMI) and adjusted mutual information (AMI). MI quantifies the information one clustering partition provides about another clustering partition; NMI normalizes MI yielding values between 0 and 1; AMI additionally

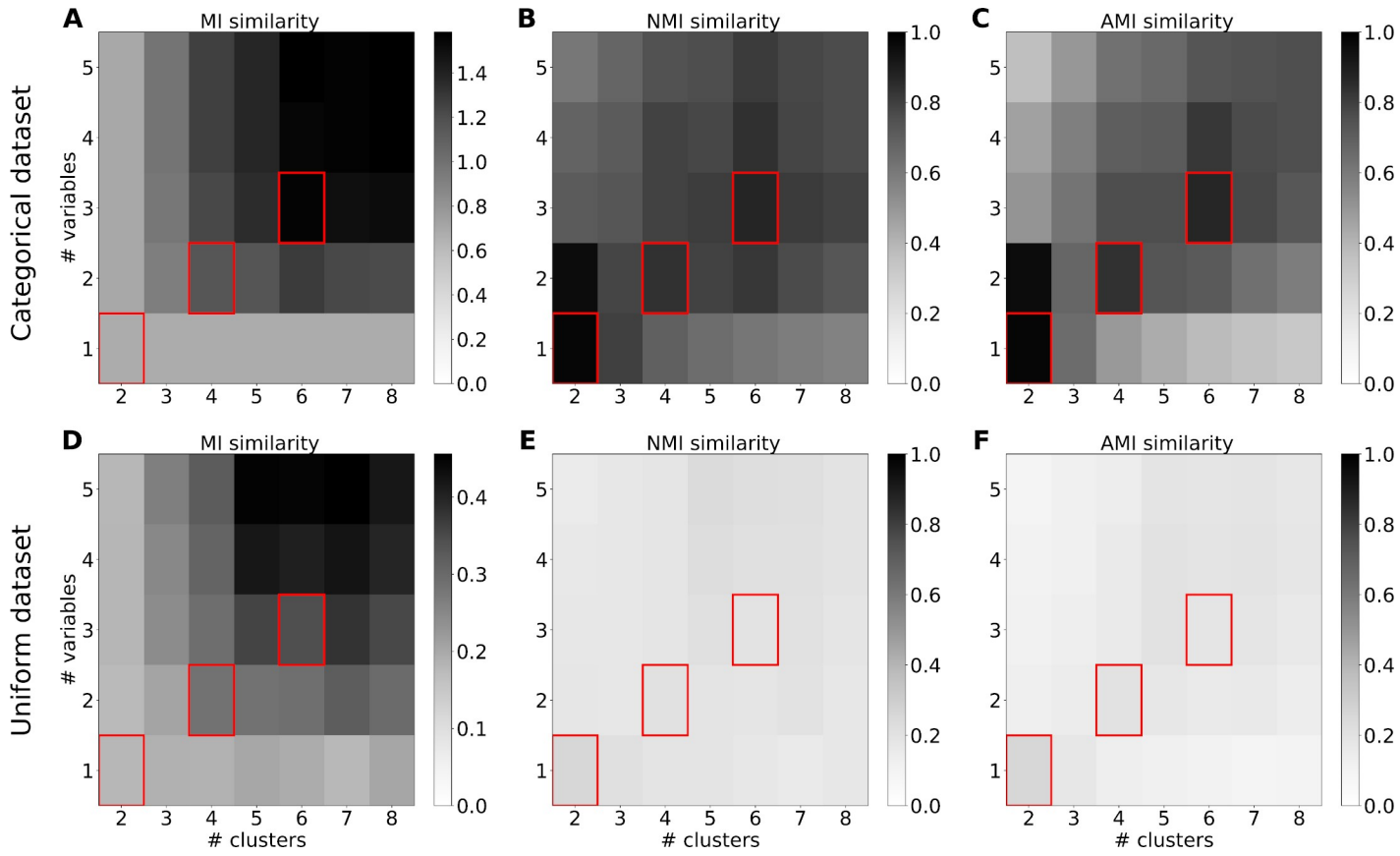

Fig 6. Comparison of different cluster similarity measures for spherical k-means partitions on synthetic data. Data were either categorical (top row) or non-categorical (bottom row) and simulated like in Figs 3 and 4. The similarity measure was either mutual information (A, D), normalized mutual information (B, E) or adjusted mutual information (C, F). The gray scale indicates the strength of similarity for given number of clusters and number of variables. Corresponding numbers of clusters and numbers of variables are marked in red.

https://doi.org/10.1371/journal.pcbi.1006667.g006

corrects for the agreement expected by chance. We compared the performance of these candidate measures of similarity using our synthetic categorical and non-categorical data sets. **Fig 6** shows the results obtained for each measure as a function of the number of clusters specified in the spherical k-means algorithm and the number of variables defined in the variable-centroid clustering. For each number $n = 1, 2, \ldots$ of variables, we tested all of the possible sets of n variables, and we identified the set providing the maximum similarity. We used exhaustive search for this purpose (see Methods).

For both categorical and non-categorical data, MI tended to increase with the number of clusters and variables (**Fig 6A and 6D**). This was expected since additional clusters and variables can convey more information about each other. Importantly, MI increased to ~0.5 bits even for non-categorical data, highlighting the necessity for normalization. The additional normalization in NMI yielded clear peaks for categorical data and mostly flat values for non-categorical data. This made it easy to discriminate between categorical and non-categorical data based on NMI. Additionally, the peaks indicated corresponding numbers of clusters and variables where n variables correspond to 2n clusters. This was because the reflection of data points on the hyper-sphere (see Methods) produced twice the number of clusters. This reflection also facilitated the separation of the data points into two clusters for both one and two variables. For this reason, the very strong peaks for two clusters should be ignored. The results

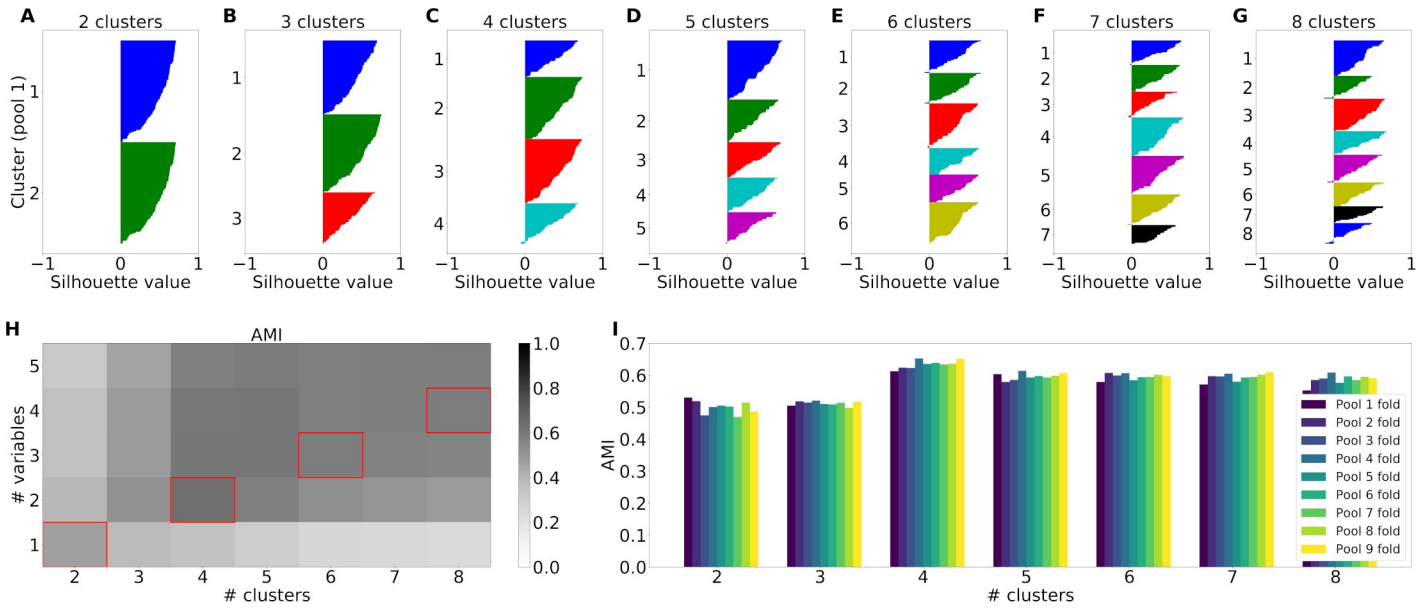

Fig 7. Cluster results for real data recorded from macaque orbitofrontal cortex limited to the post-juice time window. (A)-(G) Silhouette plots for the spherical k-means partitions of one example pool (pool 1). Each color corresponds to one cluster. The number of clusters was varied between 2 (A) and 8 (G). (H) Adjusted mutual information cluster similarity between spherical k-means clustering and variable-based centroid clustering as a function of the number of clusters and number of variables over all pools. Corresponding numbers of clusters and numbers of variables are marked in red. (I) Maximum adjusted mutual information for each number of clusters where each bar shows the result of one Jackknife fold.

https://doi.org/10.1371/journal.pcbi.1006667.g007

obtained for AMI (Fig 6C and 6F) were very similar to those for NMI. The peaks for corresponding cluster numbers and variable numbers were slightly sharper for AMI. For this reason, we selected AMI as our similarity measure for the analysis of real neural data recorded from OFC.

In conclusion, the analysis of synthetic data with known ground truth showed that a combination of spherical k-means clustering and variable-centroid clustering compared with AMI provided the most powerful approach to assess the categorical nature of neuronal representations and to identify the encoded variables.

Analysis of neuronal data

We analyzed neuronal activity recorded from OFC during experiments in which monkeys chose between different juice types (see section *Data set and previous analysis*). In total, we analyzed 9 neuronal pools, each including 139–536 neuronal responses (see Methods), where the ranges of relative juice offer values were similar within each pool (S1 Fig). Applying to each pool the same comparative clustering procedure with spherical k-means and AMI used for synthetic data, we obtained silhouette plots and a similarity profile for the neuronal data.

We varied the number of clusters between 2 and 8 and found clusters with convex silhouette plots indicating categorical data (shown for the post-juice time window in Fig 7A–7G). Moreover, the almost complete absence of negative silhouette values indicated that the spherical k-means found consistent partitions for different number of clusters. The normalized neuronal data contains 9–10 dimensions (corresponding to trial types) which are hard to visualize. In Fig 8 we illustrate the 2-dimensional projections of a data set consisting of 9-dimensional responses for the post-juice time window. Four clusters are color-coded. Even though the clusters in this representation are partly overlapping, there is a clearly discernible structure. For a qualitative assessment of the results, we examined the response prototypes defined by the

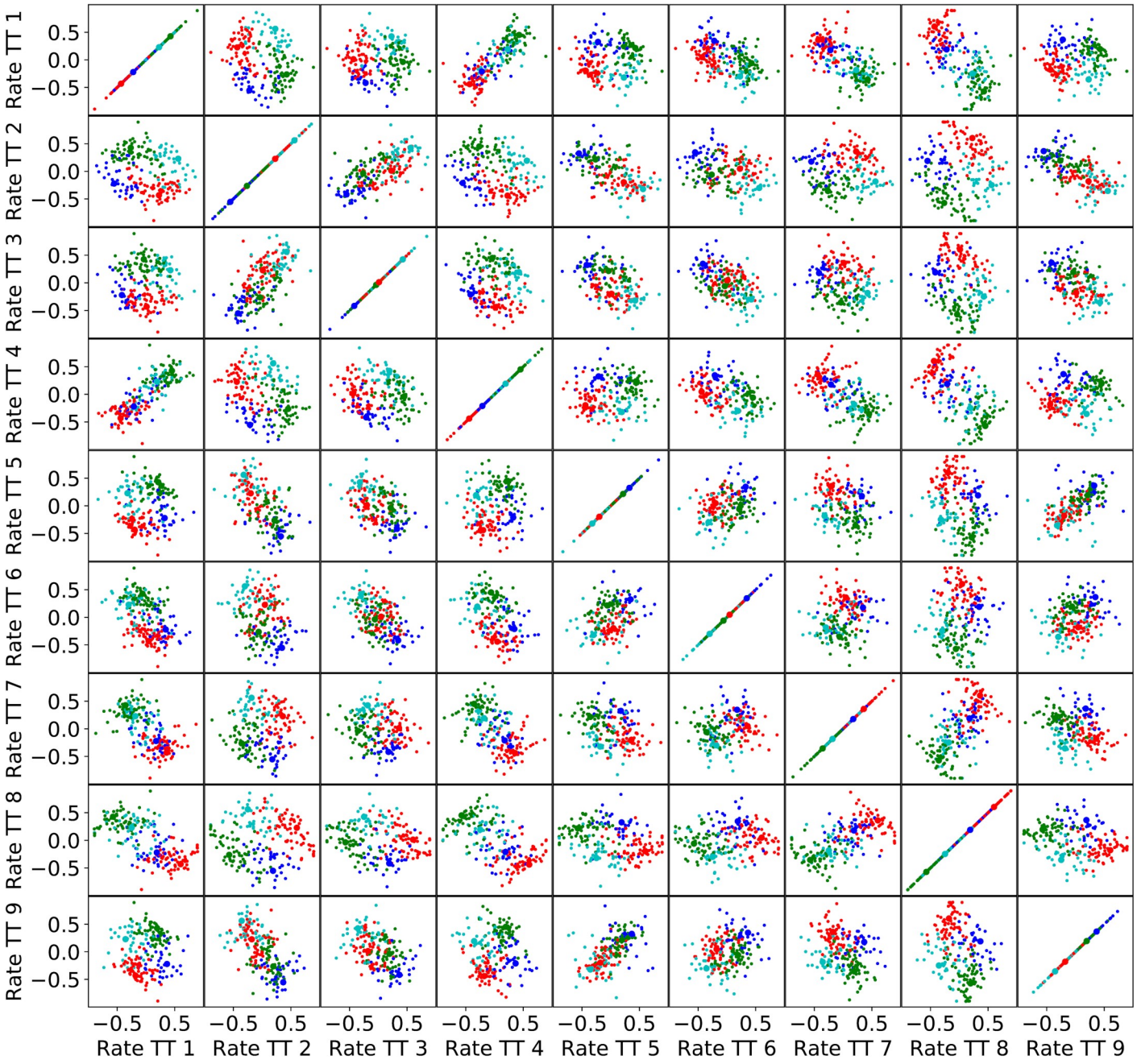

Fig 8. Visualization of four post-juice clusters in the 9-dimensional trial type space. Each color corresponds to one cluster. Each panel shows the centered and normalized firing rates of a pair of trial types and each point in a panel represents a cell from pool 1. Cluster centers are marked with black circles.

https://doi.org/10.1371/journal.pcbi.1006667.g008

centers of individual clusters. In general, the response prototypes obtained for n = 3, 4, 5 closely resembled the neuronal responses illustrated in previous studies [13, 25]. One example is illustrated in Fig 9. In other words, the clusters obtained from the spherical k-means qualitatively validated previous conclusions.

For a quantitative assessment, we used AMI. Comparing the k-means clusters and the variable-centroid clusters, we found similarity peaks for particular combinations of cluster and variable numbers (Fig 7H). These peaks resembled those obtained for synthetic data,

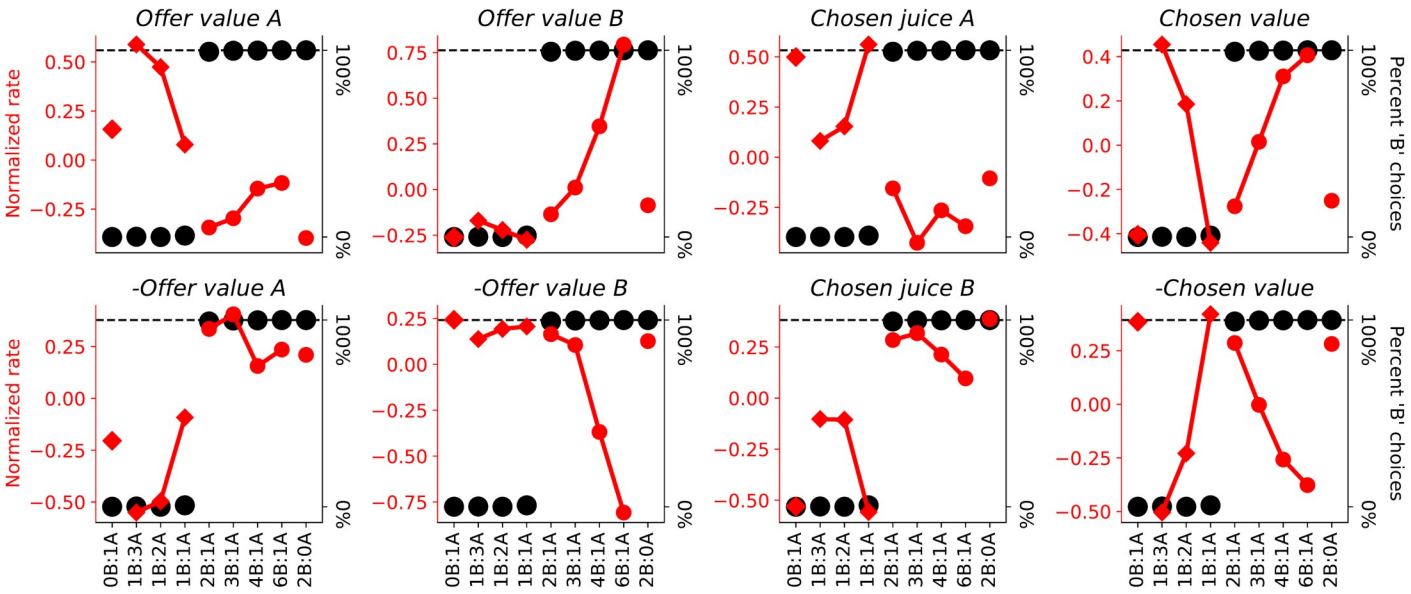

Fig 9. Tuning curves of post-juice response prototypes as defined by 8 cluster centers. The x-axis represents offer types ranked by the ratio #B:#A. The y-axis in red represents normalized response rates of pool 1. The y-axis in black shows monkey behavior. Encoded variables are denoted in the panel titles. Red diamonds represent the responses to chosen juice A whereas red dots represent the responses to chosen juice B. The separate red diamond and red dot show forced choices.

https://doi.org/10.1371/journal.pcbi.1006667.g009

providing further evidence for the categorical structure of the neural data. To analyze in more detail the clusters and variables yielding maximum AMI we performed a Jackknife analysis (see Methods). This procedure allowed us to estimate the variation of AMI values for a given number of clusters (Fig 7I). Excluding the peaks for 2 clusters, we obtained the highest AMI values for 4 clusters and 2 variables. The AMI for this combination of variables and clusters was significantly greater than the second largest AMI (Wilcoxon rank sum test, $p<0.001$). We show the corresponding tuning curves in S13 Fig. These tuning curves resemble a subset of the variables in Fig 9.

To assess the robustness of these results, we performed three control analyses: First, within each pool and each trial type, we randomly shuffled neural responses. Permuting neural responses for a given trial type destroys response patterns across trial types while preserving the distributions of responses within trial types. We expected that this would abolish the categorical representation in the data. S2 Fig shows that this was indeed the case. Of note, silhouette plots are concave (S2A–S2G Fig), resembling those that we obtained for synthetic data without a categorical representation (Fig 5A–5N). Moreover, the AMI is low (S2H Fig, S2I Fig, S3 Fig "Shuffled data"), in line with the values obtained for synthetic data without categories (Fig 6D–6F, S3 Fig "Uniform data") and unlike the data before shuffling (S3 Fig "Original data").

Second, we analyzed neural responses from the post-offer, late-delay and pre-juice time windows as well. S4 Fig, S5 Fig and S6 Fig illustrate the cluster analyses for the post-offer, late-delay and pre-juice responses, respectively. For each time window, convex silhouette plots (S4A–S4G Fig, S5A–S5G Fig, S6A–S6G Fig) and the magnitude of AMI values (S4H Fig, S4I Fig, S5H Fig, S5I Fig, S6H Fig, S6I Fig) clearly confirm the categorical nature of the representation, as did the 2-dimensional projected cluster visualizations (S7 Fig, S8 Fig, S9 Fig). For the post-offer window, a lower number of clusters yielded the highest AMI. For consistency, we show the same number of variables for the post-offer, late-delay and pre-juice time windows (S10 Fig, S11 Fig, S12 Fig, respectively). For all time windows, the response prototypes

Table 2. Selected variables.

#variables	*1*	*2*	*3*	*4*	*5*
Selected variables	*chosen value*	*chosen value A* *chosen value B*	*offer value A* *offer value B* *chosen juice*	*offer value A* *offer value B* *chosen juice* *chosen value*	*offer value A* *offer value B* *chosen juice* *chosen value* *value ratio*

https://doi.org/10.1371/journal.pcbi.1006667.t002

resembled those obtained from the post-juice data set (Fig 9, S10 Fig, S11 Fig, S12 Fig). When plotting fewer variables, we generally obtained subsets of the corresponding plots with greater number of variables (c.f. S12 Fig, Fig 9), indicating that a similar number of variables is present in the data even though expressed to varying degrees.

Third, we applied the PAIRS analysis developed by Raposo and colleagues [3] to test for the presence of neural clusters (see Methods). Confirming our results, the PAIRS analysis indicated clear categories (PAIRS index 0.67, two-sided p-value from Monte Carlo simulations < 0.001). (Of note, the number of cells included in our analysis was much larger than that in the Raposo study.)

Table 2 summarizes the results of our analyses. For 4 clusters and 2 variables, the algorithm selected variables *chosen value A* and *chosen value B* for all neuronal pools. For 6 clusters and 3 variables, variables *offer value A*, *offer value B* and *chosen juice* were selected for all pools. For 8 clusters and 4 variables, the algorithm selected variables *offer value A*, *offer value B*, *chosen value* and *chosen juice*. Note that these are the same variables identified in previous studies [13, 25]. For 10 clusters and 5 variables, the algorithm selected these same variables plus the variable *value ratio* (= other/chosen value). The substantial consistency in the variables identified with increasing numbers of clusters indicates that the results are very robust.

Discussion

We presented a new algorithm to assess whether a neuronal representation is categorical or category-free, and to identify the encoded variables if the representation is indeed categorical. The method involves two steps. First, we cluster the data without committing to any particular variable. Second, we match clusters with a set of candidate variables. Quantifying similarity between the clusters of the two steps makes it possible to identify the variables most consistent with the neuronal data. This new method overcomes limitations of previous approaches, and is widely applicable. In this study, we tested the algorithm on synthetic data and on neuronal data recorded in the primate OFC during economic decisions. With respect to the latter, the most notable result is that we found the neuronal representation in OFC to be categorical in nature. This result confirms previous assessments of this same data set [13, 25], and the results obtained by other research groups [14]. We suggest that the categorical nature of the neuronal representation sets apart OFC from other prefrontal regions, where task-relevant variables are encoded in category-free representations [6–10]. Importantly, we confirmed our result through the PAIRS analysis previously used by Raposo et al to demonstrate non-categorical encoding in the rodent posterior parietal cortex [3]. This result highlights qualitative differences between brain regions.

In addition, our algorithm identified a set of variables encoded in OFC. The variables most reliably detected–*offer value A*, *offer value B*, *chosen value* and *chosen juice*–coincide with those identified in previous studies [12, 13]. One difference concerns the number of variables. Previous work identified 4 variables imposing a criterion on the marginal explanatory power (i.e., each additional variable should explain $\geq$5% responses) [13, 25, 36]. In contrast, the AMI

criterion establishes the optimal number of variables as 2. Several elements may explain this finding. The AMI procedure penalizes the addition of further variables and thus tends to provide a conservatively small number of clusters. Concurrently, the variables encoded in OFC are substantially correlated in the experiments [13]. Geometrically, this means that the centers of different clusters are close to each other on the hyper-sphere, and not distributed randomly as might implicitly be assumed. Exacerbating this issue, in our data, neuronal responses encoding the *chosen value* have some additional jitter, because the relative value of two juices varied to some extent from session to session. This fact effectively broadened the corresponding cluster on the hyper-sphere.

Comparison with other approaches

In previous work, we assessed the categorical nature of the representation in OFC based on linear regressions and the analysis of the resulting R^2 [12, 13]. As discussed above, that approach has some limitations, addressed by the algorithm presented here. Another approach, was proposed by Raposo and colleagues [3]. The PAIRS analysis confirmed the categorical nature of the representation in OFC, but it did not identify specific variables encoded by the population.

Another clustering-based method to assess categorical encoding was recently proposed by Hirokawa et al [14]. Their data set was recorded from the rat OFC and included 42 conditions. Applying principal component analysis as a pre-processing step, they first reduced this data set to 21 dimensions. Using spectral clustering, they identified 9 clusters (the number of clusters was determined based on bootstrap stability). While there are clear similarities between their approach and ours, there are also notable differences. Both approaches are founded on clustering of pre-processed neuronal activity. Hirokawa and colleagues applied spectral clustering, while we applied spherical k-means. On simulated data, we compared silhouette plots of several clustering procedures and we found that spherical k-means performed best. Most importantly, our approach associates easily interpretable variables with the identified clusters by making use of two comparative clustering steps–spherical k-means and variable-centroid clustering. Spherical k-means operates without prior assumptions on particular variables while variable-centroid clustering can be thought of as a cluster representation of a set of variables. By selecting the set of variables most similar to the assumption-free clusters, we obtain unbiased representations of neuronal categories.

Interestingly, both our results and the results of Hirokawa et al [14] differ from those of a recent study by Blanchard et al, who concluded that the neuronal representation in OFC is category-free [15]. This apparent discrepancy highlights the advantage of assessing the categorical versus non-categorical nature of a neuronal representation without committing to any particular set of variables. Blanchard et al examined data from an experiment in which monkeys chose between two gambles. Apart from the stakes, which varied from trial to trial, the two gambles differed qualitatively–one gamble was "informative", meaning that the outcome would be revealed to the animal shortly after the choice; the other was "uninformative", meaning that the animal would learn the outcome only at the end of the trial. Informative and uninformative gambles were associated with different colors, and informativeness consistently affected choices [37]. In the analysis, the authors regressed each neuronal response recorded in OFC separately on the stakes and on the informativeness. Since the distribution of regression coefficients was not condensed along these two axes, they concluded that the representation in OFC is non-categorical [15].

A limitation of Blanchard's approach is that the neural representation may actually be categorical, but the frame of reference and/or the encoded variables may not be those tested in the analysis. Specifically, neurons in the Blanchard study might have represented the identities

and values of the offers in a color-based reference frame. Under these conditions, different groups of cells would encode the value of the informative or non-informative offers, with positive or negative sign. Such representation is categorical, but an analysis based on separate regressions on stakes and informativeness would fail to reveal its categorical nature. Similarly, an analysis of variables defined in an order-based reference frame would fail to reveal the categorical nature of the representation. To visualize this point, consider the clustering problem defined in the present study. Choosing two variables is equivalent to choosing a particular plane and to projecting all the data set from the hyper-sphere on that plane. Unless the vectors that identify the encoded variables lie on the plane, separate clusters will overlap and appear non-separable. Assessing the categorical or non-categorical nature of the representation without committing to a particular set of variables overcomes this weakness.

Categorical representation and mixed selectivity

We presented a general tool to assess whether a neuronal representation is categorical or non-categorical. Importantly, this issue is distinct from whether the encoding is pure or mixed [3]. Pure versus mixed selectivity is a property of individual cells. Consider an experiment in which conditions vary on two dimensions (e.g., visual stimuli that vary for the orientation and contrast). The activity of any given neuron could vary as a function of only one dimension (pure selectivity), as a function of a linear combination of the two dimensions (linear mixed selectivity) or as a non-linear combination of the two dimensions (non-linear mixed selectivity). In recent years, several studies have discussed the advantages of non-linear mixed selectivity [6, 9–11]. In contrast, categorical or non-categorical encoding is a property of the neuronal population [1–3]. Consider again an experiment in which conditions vary on two dimensions, referred to as variable1 and variable2. Imagine that neurons present mixed selectivity. In principle, neurons could all encode the same linear combination of the two parameters (*a1* variable + *a2* variable2, with *a1/a2* fixed for the whole population). If so, the representation would be categorical. Alternatively, different neurons could encode different linear combinations of the parameters *a1* variable + *a2* variable2, with *a1/a2* varying across the population. If so, the representation would be non-categorical. Non-categorical representations have been found in the rat posterior parietal cortex [3] and in lateral prefrontal cortex [7, 8, 38].

Non-categorical encoding implies mixed selectivity, but the converse is not true. This fact is well illustrated by the encoding of economic decision variables in OFC. By definition, subjective values integrate all the dimensions relevant to choice, including physical traits of the goods (commodity, quantity, probability, time delay, etc.) and properties internal to the subject (motivation, risk attitude, patience, etc.) [39]. For example, the subjective value of a quantity q of apple juice (A) received at time t with probability p is roughly equal (under simplifying assumptions!) to $V_A(q, p, t) \approx \rho_A\, q\, p^{\alpha}\, e^{-t/\tau}$, where ρ_A captures the subjective desirability of the apple, α captures the risk attitude, and τ captures the patience. Clearly, the value is a non-linear combination of the dimensions varied by the experimenters (q, p and t). As a consequence, any value-encoding neuron will present non-linear mixed selectivity, as indeed observed in many studies [13, 40–44]. This circumstance, however, has no implications on the categorical nature of the representation. Consider for example our task, in which monkeys choose between juice A and juice B. Variables possibly encoded by the population include *offer value A* and *offer value B*. In principle, individual neurons could encode any linear combination a_A *offer value A* + a_B *offer value B*. The categorical or non-categorical nature of the representation is a property of the joint distribution for the coefficients [a_A a_B] across the population. For example, if each neuron encodes only one of the two variables, the distribution for [a_A a_B] has two peaks at [0 1] and [1 0] and is close to zero elsewhere. Similarly, there

might be two groups of neurons encoding the value sum and the value difference. In this case, the distribution for $[a_A\ a_B]$ has two peaks centered on $[1\ 1]/2^{1/2}$ and $[1{-}1]/2^{1/2}$ and is close to zero elsewhere. In both these scenarios, the representation is categorical in nature. Conversely, coefficients $[a_A\ a_B]$ could be uniformly distributed on a broad domain, and the representation would be non-categorical. Our results demonstrate that the representation of decision variables in OFC is indeed categorical.

Methods

Experimental design and data set

The experimental procedures for data collection and preliminary data analyses have been described before [25]. Briefly, two monkeys participated in the study. All experimental procedures conformed to the NIH *Guide for the Care and Use of Laboratory Animals* and were approved by the Institutional Animal Care and Use Committee (IACUC) at Washington University in St Louis (protocol #20140031). Throughout the study, the animal health was overseen by a veterinary staff. Before training, a head restraining device and a recording chamber were implanted under general anesthesia (Isoflurane). Steps taken to increase the animal welfare included pair housing, cage enrichment, and usage of exclusively positive reinforcers.

In each session, a monkey chose between two juices (labeled A and B, with A preferred to B) offered in variable amounts. Each trial started with the animal fixating the center of a computer monitor. After 0.5 s, two sets of colored squares representing the two offers appeared on the two sides of the fixation point. For each offer, the color represented the juice type and the number of squares represented the juice amount. The animal maintained central fixation for a randomly variable delay (1–2 s), after which the fixation point was extinguished and two saccade targets appeared by the offers (go signal). The animal indicated its choice with a saccade and maintained peripheral fixation for 0.75 s before juice delivery.

In this experiment, the same neuron was recorded during two subsequent blocks of trials. Juices offered in the two blocks could be the same or different [25]. For the purpose of the present analysis, we considered data in each trial block independently. Thus each neuron appears in the analysis twice and the term "session" refers to a block of trials. In each session, offered quantities varied from trial to trial. An "offer type" was defined by two offers (e.g., [1A:3B]). Different offer types were pseudo-randomly interleaved. Their frequency varied, but each offer type was typically presented at least 20 times in each session. A "trial type" was defined by an offer and a choice (e.g., [1A:3B,A]).

In each session, choices were analyzed with a logistic regression:

$$\text{choice } B = 1/(1 + \exp(-X))$$

$$X = a_0 + a_1 \log(q_B/q_A)$$

where q_A and q_B were the quantities of juices A and B offered to the animal. The relative value of the juices was inferred from the flex of the sigmoid and defined as $\rho = \exp(-a_0/a_1)$.

Neuronal data were recorded from central OFC using standard techniques [25]. The analysis of firing rates was based on four primary time windows: post-offer (0.5 s after the offer), late delay (0.5–1 s after the offer), pre-juice (0.5 s before juice delivery), and post-juice (0.5 s after juice delivery). For each trial type and each time window, firing rates were averaged across trials. A "neuronal response" was defined as the activity of one cell in a window as a function of the trial type. Task-related responses were identified with an ANOVA (factor trial type, $p<0.001$).

In preliminary work, we submitted the present data set to standard analyses for variable selection. In these analyses, we defined a large number of variables (**Table 1**), regressed each response on each variable, and used methods for variable selection to identify a subset of variables that best explained the population (see Results and [13]). These procedures replicated previous results, as neuronal responses were found to encode variables *offer value A*, *offer value B*, *chosen value* and *chosen juice* [13].

Neuronal pools

The hyper-spherical clustering procedures introduced in this study require that different neuronal responses be defined on the same trial types (i.e., in the same space). Importantly, the offer types presented to the animal in our experiments could vary from session to session, although the same few sets of offer types were used repeatedly in many sessions. As a result, the entire data set could be divided in six groups of neuronal responses defined on the same trial types.

The variables included in the analysis are defined in **Table 1**. Of note, some variables (e.g., *chosen value*) were defined based on the relative value of the juices, which depends on the animal choices and thus varies somewhat from session to session. Ideally, the analyses described in this study would be conducted on pools of neuronal responses recorded in the same session, such that variables would be defined equally for all the responses. In contrast, our neurons were recorded in different sessions. Hence, we grouped responses in pools of similar relative values. For each group of neuronal responses recorded with the same trial types we examined the distribution of relative values. For five of the six groups, the distribution was bimodal. Hence, we split each of them in two and we removed outliers based on the inter-quartile range (IQR). In conclusion, our data set included 9 pools of neuronal responses recorded with the same trial types and similar relative values. (The remaining variability in relative values was effectively a noise factor that, if anything, made it more difficult to show categorical encoding.) Neuronal pools included 139–536 responses, and each pool was analyzed separately. When combining similarity values obtained for different pools, we weighted the similarities according to the number of neurons in the pool.

Spherical representation of neuronal responses and variables

We represented neuronal responses as points in a high-dimensional space where each axis corresponds to a trial type. Raw neuronal responses were centered (by subtracting the mean firing rate across trial types) and normalized (imposing a unitary vector length). As a result, the neuronal population was constrained to the hyper-spherical surface of unitary radius. Similarly, for each variable we calculated a vector with elements given by the variable value in each trial type. We then centered and normalized the vector. Hence, each variable was represented as a point on the unitary hyper-spherical surface.

Previous work indicated that neuronal responses can encode a variable with positive or negative slope [13]. Hence, the sign of the normalized vector is ambiguous. For this reason, before conducting the clustering procedures, we mirrored each data point on the hyper-spherical surface. Resulting cluster centers were most of the time, but not always, symmetric when adding the mirror points. In principle, non-symmetrical clusters may be understood considering even and odd numbers of clusters. Symmetry implies an even number of true clusters, which is not necessarily the case. For instance, consider the case of the 3D sphere. If the raw data present only one cluster along the equator, adding mirror points will not generate a second cluster. If we run the algorithm imposing two clusters, the algorithm will place two cluster centers somewhere on the equator, but not necessarily on opposite ends. Now consider a situation where

the raw data present a cluster along half of the equator and another cluster at one pole. Adding mirror points will result in one cluster along the equator and one cluster at each pole (3 clusters total). These examples demonstrate that mirroring does not necessarily induce an even number of clusters or symmetric cluster centers.

Variable selection procedure

We selected variables by evaluating cluster similarity of partitions induced by a set of variables and of partitions obtained from spherical k-means clustering. The general algorithm for selecting the most informative set of variables works as follows:

- For given number of clusters, partition cells using spherical *k-means* clustering yielding partition U (see below)
- For given number of variables n, select variables by:
  - Iterate over different combinations c of n variables:
    - For given set of variables c, use each variable as a cluster center and cluster cells by means of proximity clustering yielding partition V (see below)
    - Evaluate similarity between partitions U and V using adjusted mutual information (AMI)
  - Select variable combination c that maximizes similarity

We clustered cells using the spherical *k-means* algorithm [28]:

1. Start with a partitioning $\{U_i^{(0)}\}_{i=1}^R$ and the centroids $c_1^{(0)}, c_2^{(0)} \ldots, c_R^{(0)}$ associated with the partitioning. Set the index of iteration $t = 0$.
2. For each normalized rate vector x find the centroids $c_{i^*(x)}$ closest in cosine similarity to x, i.e.:

$$i*(x) = \arg\max_j \; x^T c_j^{(t)}$$

   Next, compute the new partitioning $\{U_i^{(t+1)}\}_{i=1}^R = f(\{U_i^{(t)}\}_{i=1}^R)$ induced by the old centroids $\{c_i^{(t)}\}_{i=1}^R$:

$$U_i^{(t+1)} = \{x : i*(x) = i\}, 1 \leq i \leq R$$

3. Compute new centroids corresponding to the partitioning computed for $\{U_i^{(t+1)}\}_{i=1}^R$:

$$c_i^{(t+1)} = \frac{s(U_i^{(t+1)})}{\|s(U_i^{(t+1)})\|}$$

   where $s(U_i^{(t)}) = \sum_{x \in U_i^{(t)}} x$

4. If $\sum_{i=1}^R \sum_{x \in U_i^{(t+1)}} x^T c_i - \sum_{i=1}^R \sum_{x \in U_i^{(t)}} x^T c_i$ is greater than the tolerance 1e-4 (default value), increment t by 1 and go to step 2. Otherwise, stop.

For given variables, we partitioned cells using proximity clustering: for each cell, we calculated the cosine distance to each variable and assigned the cell to the variable with the smallest distance. Variables therefore became centroids of the clusters.

Our similarity measure between partitions is the adjusted mutual information:

$$\mathrm{AMI}(U,V) = \frac{\mathrm{I}(U,V) - \mathrm{E}[\mathrm{I}(U,V)]}{\max[\mathrm{H}(U),\mathrm{H}(V)] - \mathrm{E}[\mathrm{I}(U,V)]}$$

where $U = \{U_1,\ldots,U_R\}$ and $V = \{V_1,\ldots,V_C\}$ denote two partitions of the same data (i.e. $\bigcup_i^R U_i = \bigcup_j^C V_j$). Note that a partition U here refers to a set of sets $\{U_1,\ldots,U_R\}$ where each element U_i of the partition is a set of rate vectors. E denotes expectation of the mutual information over random partitions subject to having a fixed number of clusters and points in each cluster, H denotes entropy:

$$\mathrm{H}(U) = -\sum_{i=1}^{R} \frac{a_i}{N} \log\left[\frac{a_i}{N}\right]$$

$$\mathrm{H}(V) = -\sum_{j=1}^{C} \frac{b_j}{N} \log\left[\frac{b_j}{N}\right]$$

and I(U,V) denotes mutual information [45] between U and V:

$$\mathrm{I}(U,V) = \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{n_{ij}}{N} \log\left[\frac{n_{ij}/N}{a_i b_j/N^2}\right]$$

Mutual information was used because of its several advantages as metric for computing statistical associations between neural variables or between neural and behavioral variables, namely its ability to capture all forms of associations between such variables, including both linear and non-linear ones at all orders [46]. In the above equation for I(U,V), n_{ij} denotes the number of objects that are common to clusters U_i and V_j, that is $n_{ij} = \#(U_i \cap V_j)$ and $a_i = \sum_{j=1}^{C} n_{ij}$ and $b_j = \sum_{i=1}^{R} n_{ij}$. Subtraction of the expectation values in the numerator and denominator adjusts the measure for chance and effectively corrects the positive bias of the measure. These terms can be calculated analytically [26]. In particular, we used the Python implementation *sklearn.metrics.adjusted_mutual_info_score* of the Scikit-learn package to calculate the AMI.

We checked all possible variable combinations (stopping at 5 variables) and collapsed the variables *offer value A* and *offer value B* to *offer value A|B* as well as *chosen value A* and *chosen value B* to *chosen value A|B* by pruning variable combinations that contained one but not the other of the collapsed variables. We then selected the variables and clusters with the greatest adjusted mutual information similarity.

Jackknife estimates of standard error

We estimated standard errors of adjusted mutual information values by apply the Jackknife procedure over pools [47]. For a given number of clusters and a given jackknife subsample, we took the maximum AMI over the different numbers of variables. This yielded a (#clusters x #subsamples) matrix. We then collapsed the subsample dimension in two ways: 1) For a given number of clusters, we averaged over subsamples to get the mean AMI. 2) For a given number

of clusters, we used the jackknife equation for standard deviation [47] to get an estimate of the standard error:

$$\mathrm{std}(\mathrm{AMI}) = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\theta_i - \theta_{(\cdot)})^2},$$

where θ_i denotes the i-th AMI estimate and

$$\theta_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\theta_i$$

denotes the mean AMI.

Generation of synthetic data

We generated two synthetic data sets to test our variable selection procedure: one with categories and the other without categories.

To generate the data set with categories, we selected four variables: *total value*, *offer value A*, *offer value B* and *chosen juice*. We represented each of these variables in the trial type space on the hyper-sphere as a 9-dimensional vector with unit length (see Section "Representation of cells and variables"). Then, for each of these variables, we generated 100 synthetic cell responses by adding independent Gaussian noise to each of the vector elements (zero-mean, standard deviation 0.25). Using this procedure, we obtained point clouds around each variable consisting of 100 points each. Finally, we moved the points to the unit hyper-sphere by normalizing the vector length of each point. Each point then represented the centered and normalized firing rates of a synthetic cell.

To generate a data set without categories, we drew 400 samples uniformly on the 9-dimensional unit hyper-sphere. To do so, we generated a 9-dimensional vector with independent standard normal distributed elements for each sample and then normalized the vector to unit length.

PAIRS analysis

To test whether another method for detecting the presence of neural clusters would also indicate categories, we applied the PAIRS analysis [3] as follows. For the PAIRS analysis, an input matrix is required having size (number of cells)-by-(number of trial types times number of time points per trial). For a given trial type we did not have a neural response for each cell. Therefore, we selected the nine most common trial types leaving 2380 cells that had responses for all of these trial types. We filled the matrix by calculating the 10 ms peri-stimulus time histogram for each of these cells and trial types in the time window from -500 ms before offer onset to 1,000 ms after offer onset. We then performed principal component analysis to reduce this matrix to a (number of cells)-by-8 matrix. For each cell, we then found the k nearest-neighbors and calculated the k-angle, that is the mean angle it made with each of these neighbors. The median of the angles over cells then yielded $\hat{\theta}_{\mathrm{data}}$. We then generated 10,000 matrices of size (number of cells)-by-8 filled with Gaussian random variables and calculated the k-angle for each of them. The median over the (number of cells)-times-10,000 angles then yielded $\hat{\theta}_{\mathrm{random}}$. We then calculated the PAIRS index as

$$\mathrm{PAIRS} = \frac{\hat{\theta}_{\mathrm{random}} - \hat{\theta}_{\mathrm{data}}}{\hat{\theta}_{\mathrm{random}}}.$$

We varied k between 2 and 39 and found stable PAIR indices between 0.67 and 0.70. In this range, $\hat{\theta}_{\text{random}}$ varied between 0.22 and 0.69.

Supporting information

S1 Fig. Distribution of relative values for each pool. Violin plots of the distributions of relative values. The values were log transformed for better visibility. Each component shows the distribution for one pool. The ranges of distribution support indicate that the relative values are similar in each pool.
(TIF)

S2 Fig. Shuffling analysis of cluster results for real data recorded from macaque orbitofrontal cortex. Within each pool and each trial type, neural responses were randomly shuffled to destroy categories but preserve response distributions within each trial type. (A)-(F) Silhouette plots for the spherical k-means partitions of one example pool. Each color corresponds to one cluster. The number of clusters was varied between 2 (A) and 8 (G). (H) Adjusted mutual information cluster similarity between spherical k-means clustering and variable-based centroid clustering as a function of the number of clusters and number of variables over all pools. Corresponding numbers of clusters and numbers of variables are marked in red. (I) Maximum adjusted mutual information for each number of clusters with Jackknife estimated standard errors.
(TIF)

S3 Fig. Comparison of maximum adjusted mutual information for non-categorical data, shuffled data and unshuffled data. Maximum adjusted mutual information over number of variables for different numbers of clusters as in Fig 6F ("Uniform data"), S2 Fig I collapsed over folds ("Shuffled data") and Fig 7I collapsed over folds ("Original data").
(TIF)

S4 Fig. Cluster results for real data limited to the post-offer time window. In this analysis, neural responses were taken from the post-offer time window only. (A)-(G) Silhouette plots for the spherical k-means partitions of one example pool. Each color corresponds to one cluster. The number of clusters was varied between 2 (A) and 8 (G). (H) Adjusted mutual information cluster similarity between spherical k-means clustering and variable-based centroid clustering as a function of the number of clusters and number of variables over all pools. Corresponding numbers of clusters and numbers of variables are marked in red. (I) Maximum adjusted mutual information for each number of clusters where each bar shows the result of one Jackknife fold.
(TIF)

S5 Fig. Cluster results for real data limited to the late-delay time window. In this analysis, neural responses were taken from the late-delay time window only. (A)-(G) Silhouette plots for the spherical k-means partitions of one example pool. Each color corresponds to one cluster. The number of clusters was varied between 2 (A) and 8 (G). (H) Adjusted mutual information cluster similarity between spherical k-means clustering and variable-based centroid clustering as a function of the number of clusters and number of variables over all pools. Corresponding numbers of clusters and numbers of variables are marked in red. (I) Maximum adjusted mutual information for each number of clusters where each bar shows the result of one Jackknife fold.
(TIF)

S6 Fig. Cluster results for real data limited to the pre-juice time window. In this analysis, neural responses were taken from the pre-juice time window only. (A)-(G) Silhouette plots for the spherical k-means partitions of one example pool. Each color corresponds to one cluster. The number of clusters was varied between 2 (A) and 8 (G). (H) Adjusted mutual information cluster similarity between spherical k-means clustering and variable-based centroid clustering as a function of the number of clusters and number of variables over all pools. Corresponding numbers of clusters and numbers of variables are marked in red. (I) Maximum adjusted mutual information for each number of clusters where each bar shows the result of one Jack-knife fold.
(TIF)

S7 Fig. Visualization of four post-offer clusters in the 9-dimensional trial type space. Like Fig 8 but neural responses were taken from the post-offer time window only. Each color corresponds to one cluster. Each panel shows the centered and normalized firing rates of a pair of trial types and each point in a panel represents a cell from pool 1. Cluster centers are marked with black circles.
(TIF)

S8 Fig. Visualization of four late-delay clusters in the 9-dimensional trial type space. Like Fig 8 but neural responses were taken from the late-delay time window only. Each color corresponds to one cluster. Each panel shows the centered and normalized firing rates of a pair of trial types and each point in a panel represents a cell from pool 1. Cluster centers are marked with black circles.
(TIF)

S9 Fig. Visualization of four pre-juice clusters in the 9-dimensional trial type space. Like Fig 8 but neural responses were taken from the pre-juice time window only. Each color corresponds to one cluster. Each panel shows the centered and normalized firing rates of a pair of trial types and each point in a panel represents a cell from pool 1. Cluster centers are marked with black circles.
(TIF)

S10 Fig. Tuning curves of post-offer response prototypes as defined by 6 cluster centers. Like Fig 9 but neural responses were taken from the post-offer time window only. The number of clusters is lower than for the full data set (6 clusters instead of 8 clusters). The x-axis represents offer types ranked by the ratio #B:#A. The y-axis in red represents normalized response rates. The y-axis in black shows monkey behavior. Red diamonds represent the responses to chosen juice A whereas red dots represent the responses to chosen juice B. The separate red diamond and red dot show forced choices.
(TIF)

S11 Fig. Tuning curves of late-delay response prototypes as defined by 8 cluster centers. Like Fig 9 but neural responses were taken from the late-delay time window only. The number of clusters is the same as for the full data set. The x-axis represents offer types ranked by the ratio #B:#A. The y-axis in red represents normalized response rates. The y-axis in black shows monkey behavior. Red diamonds represent the responses to chosen juice A whereas red dots represent the responses to chosen juice B. The separate red diamond and red dot show forced choices.
(TIF)

S12 Fig. Tuning curves of pre-juice response prototypes as defined by 8 cluster centers. Like Fig 9 but neural responses were taken from the pre-juice time window only. The number

of clusters is the same as for the full data set. The x-axis represents offer types ranked by the ratio #B:#A. The y-axis in red represents normalized response rates. The y-axis in black shows monkey behavior. Red diamonds represent the responses to chosen juice A whereas red dots represent the responses to chosen juice B. The separate red diamond and red dot show forced choices.
(TIF)

S13 Fig. Tuning curves of post-juice response prototypes as defined by 4 cluster centers. Like Fig 9 but for 4 cluster and 2 variables. The x-axis represents offer types ranked by the ratio #B:#A. The y-axis in red represents normalized response rates. The y-axis in black shows monkey behavior. Red diamonds represent the responses to chosen juice A whereas red dots represent the responses to chosen juice B. The separate red diamond and red dot show forced choices.
(TIF)

Author Contributions

Conceptualization: Stefano Panzeri, Camillo Padoa-Schioppa.

Data curation: Jue Xie, Camillo Padoa-Schioppa.

Funding acquisition: Arno Onken, Stefano Panzeri, Camillo Padoa-Schioppa.

Investigation: Arno Onken, Stefano Panzeri, Camillo Padoa-Schioppa.

Methodology: Arno Onken, Stefano Panzeri, Camillo Padoa-Schioppa.

Project administration: Stefano Panzeri, Camillo Padoa-Schioppa.

Resources: Stefano Panzeri.

Software: Arno Onken.

Supervision: Stefano Panzeri, Camillo Padoa-Schioppa.

Validation: Arno Onken, Stefano Panzeri, Camillo Padoa-Schioppa.

Visualization: Arno Onken.

Writing – original draft: Arno Onken, Stefano Panzeri, Camillo Padoa-Schioppa.

Writing – review & editing: Arno Onken, Stefano Panzeri, Camillo Padoa-Schioppa.

References

1. Pouget A, Sejnowski TJ. Spatial transformations in the parietal cortex using basis functions. J Cogn Neurosci. 1997; 9(2):222–37. https://doi.org/10.1162/jocn.1997.9.2.222 PMID: 23962013
2. Salinas E. Context-dependent selection of visuomotor maps. BMC neuroscience. 2004; 5:47. https://doi.org/10.1186/1471-2202-5-47 PMID: 15563737
3. Raposo D, Kaufman MT, Churchland AK. A category-free neural population supports evolving demands during decision-making. Nat Neurosci. 2014; 17(12):1784–92. https://doi.org/10.1038/nn.3865 PMID: 25383902
4. Wallis JD. Cross-species studies of orbitofrontal cortex and value-based decision-making. Nat Neurosci. 2012; 15(1):13–9. https://doi.org/10.1038/nn.2956 PMID: 22101646
5. Padoa-Schioppa C, Conen KE. Orbitofrontal cortex: A neural circuit for economic decisions. Neuron. 2017; 96:736–54. https://doi.org/10.1016/j.neuron.2017.09.031 PMID: 29144973
6. Rigotti M, Barak O, Warden MR, Wang XJ, Daw ND, Miller EK, et al. The importance of mixed selectivity in complex cognitive tasks. Nature. 2013; 497(7451):585–90. https://doi.org/10.1038/nature12160 PMID: 23685452

7. Mante V, Sussillo D, Shenoy KV, Newsome WT. Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature. 2013; 503(7474):78–84. https://doi.org/10.1038/nature12742 PMID: 24201281
8. Cai X, Padoa-Schioppa C. Contributions of orbitofrontal and lateral prefrontal cortices to economic choice and the good-to-action transformation. Neuron. 2014; 81(5):1140–51. S0896-6273(14)00013-0 [pii] https://doi.org/10.1016/j.neuron.2014.01.008 PMID: 24529981
9. Enel P, Procyk E, Quilodran R, Dominey PF. Reservoir computing properties of neural dynamics in prefrontal cortex. PLoS Comput Biol. 2016; 12(6):e1004967. https://doi.org/10.1371/journal.pcbi.1004967 PMID: 27286251
10. Parthasarathy A, Herikstad R, Bong JH, Medina FS, Libedinsky C, Yen SC. Mixed selectivity morphs population codes in prefrontal cortex. Nat Neurosci. 2017; 20(12):1770–9. https://doi.org/10.1038/s41593-017-0003-2 PMID: 29184197
11. Fusi S, Miller EK, Rigotti M. Why neurons mix: high dimensionality for higher cognition. Curr Opin Neurobiol. 2016; 37:66–74. https://doi.org/10.1016/j.conb.2016.01.010 PMID: 26851755
12. Padoa-Schioppa C. Neuronal origins of choice variability in economic decisions. Neuron. 2013; 80 (5):1322–36. S0896-6273(13)00844-1 [pii] https://doi.org/10.1016/j.neuron.2013.09.013 PMID: 24314733
13. Padoa-Schioppa C, Assad JA. Neurons in orbitofrontal cortex encode economic value. Nature. 2006; 441(7090):223–6. https://doi.org/10.1038/nature04676 PMID: 16633341
14. Hirokawa J, Vaughan A, Kepecs A. Categorical representations of decision-variables in orbitofrontal cortex. BioRxiv. 2017.
15. Blanchard TC, Piantadosi ST, Hayden BY. Robust mixture modeling reveals category-free selectivity in reward region neuronal ensembles. J Neurophysiol. 2018; 119(4):1305–18. https://doi.org/10.1152/jn.00808.2017 PMID: 29212924
16. Hunt LT, Kolling N, Soltani A, Woolrich MW, Rushworth MFS, Behrens TEJ. Mechanisms underlying cortical activity during value-guided choice. Nat Neurosci. 2012; 15(3):470–U169. https://doi.org/10.1038/nn.3017 PMID: 22231429
17. Rustichini A, Padoa-Schioppa C. A neuro-computational model of economic decisions. J Neurophysiol. 2015; 114(3):1382–98. https://doi.org/10.1152/jn.00184.2015 PMID: 26063776
18. Solway A, Botvinick MM. Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates. Psychological Review. 2012; 119(1):120–54. https://doi.org/10.1037/a0026435 PMID: 22229491
19. Friedrich J, Lengyel M. Goal-directed decision making with spiking neurons. J Neurosci. 2016; 36 (5):1529–46. https://doi.org/10.1523/JNEUROSCI.2854-15.2016 PMID: 26843636
20. Song HF, Yang GR, Wang XJ. Reward-based training of recurrent neural networks for cognitive and value-based tasks. Elife [Internet]. 2017 Jan 13 5293493]; 6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28084991.
21. Louie K, LoFaro T, Webb R, Glimcher PW. Dynamic divisive normalization predicts time-varying value coding in decision-related circuits. J Neurosci. 2014; 34(48):16046–57. https://doi.org/10.1523/JNEUROSCI.2851-14.2014 PMID: 25429145
22. Cisek P. Making decisions through a distributed consensus. Curr Opin Neurobiol. 2012; 22(6):927–36. https://doi.org/10.1016/j.conb.2012.05.007 PMID: 22683275
23. Hare TA, Schultz W, Camerer CF, O'Doherty JP, Rangel A. Transformation of stimulus value signals into motor commands during simple choice. P Natl Acad Sci USA. 2011; 108(44):18120–5. https://doi.org/10.1073/pnas.1109322108
24. Zhong S, editor Efficient online spherical k-means clustering. Neural Networks, 2005 IJCNN'05 Proceedings 2005 IEEE International Joint Conference on; 2005: IEEE.
25. Xie J, Padoa-Schioppa C. Neuronal remapping and circuit persistence in economic decisions. Nat Neurosci. 2016; 19(6):855–61. https://doi.org/10.1038/nn.4300 PMID: 27159800
26. Vinh NX, Epps J, Bailey J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. J Mach Learn Res. 2010; 11:2837–54.
27. Sculley D. Web-scale k-means clustering. Proceedings of the 19th international conference on World wide web; Raleigh, North Carolina, USA. 1772862: ACM; 2010. p. 1177–8.
28. Dhillon IS, Guan Y, Kogan J. Iterative clustering of high dimensional text data augmented by local search. 2002 IEEE International Conference on Data Mining, 2002 Proceedings. 2002:131–8. https://doi.org/10.1109/ICDM.2002.1183895
29. Ward JH Jr. Hierarchical grouping to optimize an objective function. J Am Stat Assoc. 1963; 58:236–44.

30. Rokach L, Maimon O. Clustering methods. Data mining and knowledge discovery handbook. New York: Springer; 2005. p. 321–52.
31. Zhang T, Ramakrishnan R, Livny M. BIRCH: An efficient data clustering method for very large databases. SIGMOD Rec. 1996; 25(2):103–14. https://doi.org/10.1145/235968.233324
32. Yu SX, Shi J. Multiclass spectral clustering. Proceedings of the Ninth IEEE International Conference on Computer Vision—Volume 2. 946658: IEEE Computer Society; 2003. p. 313.
33. Rousseeuw P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987; 20(1):53–65. https://doi.org/10.1016/0377-0427(87)90125-7
34. Hubert L, Arabie P. Comparing partitions. J Classif. 1985; 2:193–218.
35. Rosenberg A, Hirschberg J. V-measure: A conditional entropy-based external cluster evaluation measure. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). 2007:410–20.
36. Padoa-Schioppa C, Assad JA. The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. Nat Neurosci. 2008; 11(1):95–102. https://doi.org/10.1038/nn2020 PMID: 18066060
37. Blanchard TC, Hayden BY, Bromberg-Martin ES. Orbitofrontal cortex uses distinct codes for different choice attributes in decisions motivated by curiosity. Neuron. 2015; 85(3):602–14. https://doi.org/10.1016/j.neuron.2014.12.050 PMID: 25619657
38. Wallis JD, Anderson KC, Miller EK. Single neurons in prefrontal cortex encode abstract rules. Nature. 2001; 411(6840):953–6. https://doi.org/10.1038/35082081 PMID: 11418860
39. Padoa-Schioppa C. Neurobiology of economic choice: a good-based model. Annu Rev Neurosci. 2011; 34:333–59. https://doi.org/10.1146/annurev-neuro-061010-113648 PMID: 21456961
40. Saez A, Rigotti M, Ostojic S, Fusi S, Salzman CD. Abstract context representations in primate amygdala and prefrontal cortex. Neuron. 2015; 87(4):869–81. https://doi.org/10.1016/j.neuron.2015.07.024 PMID: 26291167
41. Raghuraman AP, Padoa-Schioppa C. Integration of multiple determinants in the neuronal computation of economic values. J Neurosci. 2014; 34(35):11583–603. 34/35/11583 [pii] https://doi.org/10.1523/JNEUROSCI.1235-14.2014 PMID: 25164656
42. O'Neill M, Schultz W. Coding of reward risk by orbitofrontal neurons is mostly distinct from coding of reward value. Neuron. 2010; 68(4):789–800. S0896-6273(10)00773-7 [pii] https://doi.org/10.1016/j.neuron.2010.09.031 PMID: 21092866
43. Cai X, Kim S, Lee D. Heterogeneous coding of temporally discounted values in the dorsal and ventral striatum during intertemporal choice. Neuron. 2011; 69(1):170–82. S0896-6273(10)00983-9 [pii] https://doi.org/10.1016/j.neuron.2010.11.041 PMID: 21220107
44. Louie K, Glimcher PW. Separating value from choice: delay discounting activity in the lateral intraparietal area. J Neurosci. 2010; 30(16):5498–507. 30/16/5498 [pii] https://doi.org/10.1523/JNEUROSCI.5742-09.2010 PMID: 20410103
45. Shannon CE. A mathematical theory of communication. Bell Sys Tech J. 1948; 27:379–423.
46. Quian Quiroga R, Panzeri S. Extracting information from neuronal populations: information theory and decoding approaches. Nat Rev Neurosci. 2009; 10(3):173–85. https://doi.org/10.1038/nrn2578 PMID: 19229240
47. Efron B, Stein C. The Jackknife estimate of variance. Ann Stat. 1981; 9(3):586–96.